CSCE 421: Spring 2023 Homework 1

Assigned Jan 17, due on Mon, Feb 2, by 11:59 PM.

Submit your assignments (written+coding) seperately on gradescope. Please name your coding assignment as 'assignment1.py'. Use the provided python template file, complete ONLY the functions (DO NOT edit function definitions, code outside the function, or use any other libraries).

A Few Notes:

- Coding assignments should be done only in Python.
- Please start early! This includes learning how to use Latex!
- For solution please use the following template `https://www.overleaf.com/latex/templates/neurips-2022/kxymzbjpwsqx`.
- if you need to use a different editor to write the solutions - please contact the TA and Instructor first
- This is an individual assignment. While you are welcome to discuss general concepts together and on the discussion board your solutions must be yours and yours alone.
- **SHOW YOUR WORK.**

Problem 1: Gradient Calculation. NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

In this question you are required to calculate gradients for 2 scalar functions.

(1) Calculate the gradient of the function $f(x, y) = x^2 + ln(y) + xy + y^3$. What is the gradient value for $(x, y) = (10, -10)$?

(2) Calculate the gradient of the function $f(x, y, z) = tanh(x^3y^3) + sin(z^2)$. What is the gradient value for $(x, y, z) = (-1, 0, \pi/2)$?

Problem 2: Matrix Multiplication. NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

In this question you are required to perform matrix multiplication.

(1)

$$\begin{bmatrix} 10 \\ -5 \\ 2 \\ 8 \end{bmatrix} \begin{bmatrix} 0 & 3 & 0 & 1 \end{bmatrix} = ?$$

(2)

$$\begin{bmatrix} 7 & -3 & 1 & 9 \end{bmatrix} \begin{bmatrix} -3 \\ -4 \\ 6 \\ 0 \end{bmatrix} = ?$$

(3)

$$\begin{bmatrix} 1 & -1 & 6 & 7 \\ 9 & 0 & 8 & 1 \\ -8 & 1 & 2 & 3 \\ 10 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 6 & 2 & 0 \\ 0 & -1 & 1 \\ -3 & 0 & 4 \\ 3 & 4 & 7 \end{bmatrix} =?$$

Problem 3: Vector Norms. NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

Consider these two points in the 3-dimensional space:

$$\mathbf{a} = \begin{bmatrix} 7 \\ 0 \\ -1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 7 \\ 9 \\ -5 \end{bmatrix}$$

Calculate their distance using the following norms:

(1) ℓ_0

(2) ℓ_1

(3) ℓ_2

(4) ℓ_∞

Problem 4: Probability Calculation. NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

Consider a problem where we are rolling 2 dices where each dice has 6 faces numbered from 1 to 6. Answer the following questions:

(1) What is the sample space?

(2) If the event we are interested in is the sum being 10, what would be the probability of observing such an event?

(3) If the event we are interested in is the sum being 6, what would be the probability of observing such an event?

Problem 5: Mean/Variance Calculation. NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

Assume we have a random variable X with a Uniform probability density function. Uniform probability density is defined as:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & otherwise \end{cases}$$

(1) What is the mean of X?

(2) What is the standard deviation of X?

Problem 6: Classification Quality Metric Computation. NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

Assume you take an avocado detector to your favorite restaurant to detect tacos with no avocados inside, because they have been forgetting of late and never let you replace them once it is handed to you. Here is the confusion matrix of the avocado detector:

		ground truth	
		avocado	no avocado
Avocado detector	avocado	37	23
	no avocado	45	55

(1) What is the accuracy of the detector?

(2) What is the balanced accuracy of the detector?

(3) What is the precision of the detector?

(4) What is the recall of the detector?

(5) What is the F1-measure of the detector?

Problem 7: ROC Computation. NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.

In Problem 6, assume that their microwave avocado detector does not give a binary output regarding the existence of avocados inside the taco. Alternatively, it outputs a probability of such an event. Jose, a CS sophomore who wants to put his knowledge to practice, wants to approximate the AUROC of the detector using 5 points as candidate thresholds: $\{0, 0.25, 0.5, 0.75, 1\}$. In a few tests that they ran, the probabilities and their corresponding ground truths were as follows:

predicted	ground truth
10%	0
5%	0
70%	1
50%	0
90%	1
65%	1
35%	1
60%	0
15%	1
20%	0

Note: If threshold is 0.5 and prediction is 50% then it's false. Please help him by computing the following:

(1) What would be the ROC values (TPR/FPR) for threshold = 0?

(2) What would be the ROC values (TPR/FPR) for threshold = 0.25?

(3) What would be the ROC values (TPR/FPR) for threshold = 0.5?

(4) What would be the ROC values (TPR/FPR) for threshold = 0.75?

(5) What would be the ROC values (TPR/FPR) for threshold = 1?

(6) What would be the AUROC approximation using the above results? (HINT: remember Riemann sum)

Problem 8: Coding K-NN.

This is a coding assignment. Throughout the course, you will have several coding assignments. You are required to use python for this course.

In this assignment we will implement k-NN. More specifically, we are interested in seeing the effect of varying k on the performance.

The dataset we will use in this assignment is named *Smarket* (can be downloaded from here) `https://github.com/jcrouser/islr-python/blob/master/data/Smarket.csv`.

Your submission should follow instructions stated at the beginning of assignment and should have full points on gradescope. Any submission with an error would result in 0 points. Unless stated otherwise, use default parameters provided by the libraries whenever in confusion.

(1) Fill in the function *read_data*, that takes in the filename as string, and returns a pandas dataframe. Hint: you may find *read csv* function from *pandas* library useful

(2) Fill in the function *get_df_shape*, that takes in the pandas dataframe as input and returns the shape as a tuple as output. Shape means the dimensions of the data. Hint: *pandas* dataframe instances have a variable *shape*

(3) Fill in the function *extract_features_label* that returns features and the label from the data. See the function definition (and sample main function) for input/output types. The features we are interested in are *Lag1* and *Lag2* and the label is *Direction*.

(4) Fill in the function *data_split* that given features and labels split the data into a train/test split (with a given ratio). The function returns 4 numpy arrays in the following order *x_train, y_train, x_test, y_test*. Hint: you can use *train test split* from *sklearn* library. Do not shuffle the data for the assignment.

(5) Fill in the function *knn_test_score* that takes the train test data as generated by previous question and applies kNNs on the train data, and retuns accuracy on the test data given the number of neighbours to use as first parameters. Hint: You can use the *KNeighborsClassifier* function from *sklearn* library.

(6) Fill in the function *knn_evaluate_with_neighbours* that takes the data (same as above) but iterates over (n_neighbours_min to n_neighbours_max). Note, both min and max are inclusive.